ISSN 2005-8063
2016. 6. 30.
Vol.8 No.2
pp. 11-16

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

http://dx.doi.org/10.13064/KSSS.2016.8.2.011

# Phoneme distribution and syllable structure of entry words in the CMU English Pronouncing Dictionary

Byunggon Yang*

**Abstract**

This study explores the phoneme distribution and syllable structure of entry words in the CMU English Pronouncing Dictionary to provide phoneticians and linguists with fundamental phonetic data on English word components. Entry words in the dictionary file were syllabified using an R script and examined to obtain the following results: First, English words preferred consonants to vowels in their word components. In addition, monophthongs occurred much more frequently than diphthongs. When all consonants were categorized by manner and place, the distribution indicated the frequency order of stops, fricatives, and nasals according to manner and that of alveolars, bilabials and velars according to place. These results were comparable to the results obtained from the Buckeye Corpus (Yang, 2012). Second, from the analysis of syllable structure, two-syllable words were most favored, followed by three- and one-syllable words. Of the words in the dictionary, 92.7% consisted of one, two or three syllables. This result may be related to human memory or decoding time. Third, the English words tended to exhibit discord between onset and coda consonants and between adjacent vowels. Dissimilarity between the last onset and the first coda was found in 93.3% of the syllables, while 91.6% of the adjacent vowels were different. From the results above, the author concludes that an analysis of the phonetic symbols in a dictionary may lead to a deeper understanding of English word structures and components.

**Keywords:** CMU English Pronouncing Dictionary, phoneme distribution, syllable structure, discord

## 1. Introduction

English words consist of syllables, which can be decomposed into smaller sound units of consonants and vowels. The syllables and those consonant and vowel phonemes are the building blocks of an English word. In the past, only a few limited analyses of general syllable structures and components were possible because of complicated syllabification procedures and the tremendous time needed to summarize patterns of phonetic symbols by breaking down an enormous number of words into their components. Recently, publicly available computer analysis software has provided solutions for handling big data and delving into the unknown big picture behind the data.

Native English speakers pronounce a word such as “understanding” not as a simple string of sounds but as a group of sounds that make beats based on the four vowels, as in [ʌn · də· · stæn · diŋ] (Cable, 2013). Those beats constitute the syllables. Every syllable consists of an onset, nucleus, and coda. Williamson (2014) described the English syllable structure as a nucleus containing only one vowel, either a monophthong or diphthong, with up to three consonants placed before it as the onset. Williamson (2014, Figures 1 & 2) diagrammatically presented 26 two-consonant onset clusters and 6 three-consonant clusters adapted from Jackson (1980). Duanmu (1997:13) had already compiled an exhaustive list of 56 possible onset clusters. However, diagramming the coda was more difficult because of its complexity; Williamson (2014) mentioned at least 48 allowable three-consonant clusters and seven allowable four-consonant clusters.

* English Education Dept., Pusan National University, bgyang@pusan.ac.kr
Received 1 May 2016; Revised 7 June 2016; Accepted 18 June 2016

The syllabification algorithm identifies vowels in a given entry word and then assigns all permissible onsets to a given syllable and then to the codas. Using the syllabified data, we can examine the syllable structure and the distribution of phonemes. There have been a few attempts to analyze English syllable characteristics (Duanmu, 2002; Goldsmith, 1990; Kessler & Treiman, 1997; McMahon, 2002). Phonotactic constraints refer to the restrictions that determine which onsets or codas are possible (McMahon, 2002). For example, the first consonant of a CCC onset must be /s/, and coda clusters of nasal plus oral stop are acceptable only if the two stops share the same place of articulation (McMahon, 2002:106). Generally, the nucleus has the highest sonority, with the sonority of onsets or codas slowly decreasing before and after the nucleus. Goldsmith (1990) characterized the English syllable as having a particular type of internal structure. Some linguists, such as Davis (1985), rejected all arguments for an internal syllable structure by pointing to the exceptions. However, the author claims that taking a closer look at the general patterns of syllables may enable us to identify universal properties of language.

Duanmu (2002) noted that the sonority-based theory of English syllables cannot explain fully permissible and non-permissible onset clusters. The theory requires that the first sound of the onset cluster be less sonorous than the second sound, as observed in the onset clusters of the English words “bring” and “flow”. Such clusters as [pl] and [tl] are considered to have the same sonority slope, but the former is permissible while the latter is not. An additional place dissimilation constraint may solve the issue, but it is not applicable to [dr] and [tr]. Thus, Duanmu (2002) proposed an alternative articulator-based feature theory. Here, the articulators are defined as the movable physiological organs involved in speech production, and the feature indicates a gesture of the articulator. Instead of considering the onset clusters to be a combination of two different sounds, Duanmu (2002) appropriated a single onset or coda slot that can be filled with a single or complex sound defined by both specific articulators and features excluding initial coronal fricatives.

In another study, Kessler & Treiman (1997) analyzed 2001 monomorphemic CVC words in the unabridged Random House Dictionary (Flexner, 1987). They examined only Anglicized words, screening out words with foreign phonemes or accented letters, foreign measures, or place and ethnic names. Their results showed that coronal consonants preferred the coda position significantly more than non-coronals did. Among coronals, anterior consonants had a more marked tendency to appear in the coda than do non-anterior coronals (Kessler & Treiman, 1997: 301). Those authors reported that /d/ favored the onset position more than other anterior coronals did. In addition, /z, ɵ, n, t, l, k/ showed a significant preference for the onset position, while /b, ǰ, ʃ, r/ tended toward the coda position. Those authors also investigated association patterns among the onset, vowel and coda and found that the vowel-coda association was always stronger than the onset-coda association. Moreover, they described tendencies toward onset-coda dissimilarity in English CVC words. Among onset-coda pairings, all patterns pointed to favoring discords in terms of both the manner and place of articulation. For example, if the onsets were coronal, the codas were non-coronal, and the reverse was also true. Research has also noted the isochrony in the syllable structure (Berg, 1994): phonetically longer vowels tend to pair with phonetically shorter consonants, such as coronals (Crystal & House, 1988). In analyses of syllable structures, Borowsky (1989) and Rubach & Booij (1990) reported that English and Polish words have complicated onset and coda clusters primarily at the beginning and end of words but not in the middle of a morpheme.

From the studies described above, one could easily note that general syllable patterns can be traced through a quantitative analysis of English words. Thus far, few articles on this issue have been published because of the demanding manual or computation procedures. This study attempts to extend that previous research with much more phonetic data from a currently available pronunciation dictionary. However, because of journal space limitations, this study will not present extensive, detailed results on syllable structures and components.

The objectives of this paper are twofold: to examine the phoneme distribution of English words and to phonotactically analyze the syllable structures and components of English words. Specifically, the pronunciation symbols of the entry words in an online pronunciation dictionary will be classified by syllables using R. Then, the phoneme distributions in the syllable data and specific patterns of syllable components will be examined at both the segmental and categorical levels. Finally, adjacent syllable components will be compared to examine any prevailing pattern among them. The results may enhance our understanding of English syllables and may be applicable to speech recognition based on the distribution pattern of English segments in syllables.

## 2. Method

### 2.1. CMU English Pronouncing Dictionary

This study examined English syllable structure from the pronunciation symbols used for the entry words in the Carnegie Mellon University Pronouncing Dictionary, which is an open-source, machine-readable pronunciation dictionary for North American English that contains over 134,000 entry words and their phonetic symbols and is available online (http://www.speech.cs.cmu.edu/cgi-bin/cmudict). The author chose this dictionary because it lists almost all of the English words currently used in North America and thus may best represent English word structures and components. The downloaded file has 126 lines of various information at the beginning of the dictionary that were deleted, along with the last 5 lines from the 133780th entry word. Moreover, several entry words that included numbers (c1, m1) or other miscellaneous symbols ({, ‘ , -) were also deleted to establish the final data set of 116588 entry words using the internal functions of Microsoft Excel. The shortest word consists of one alphabet character such as “a” or “b”, while the longest word has 34 characters with 14 syllables. Because of its enormous size, the author of the current study did not apply any further screening as Kessler & Treiman (1997) manually did with the 2001 non-anglicized words in their study.

### 2.2. Syllabification Procedure

Words in the CMU dictionary were processed using the basic syllabification rules prescribed in Noyer (2016). First, the nucleus vowel of a given word was assigned, and single or complex onsets were adjoined to the given vowel. Such on-glide after tautosyllabic consonants in a CGV syllable type was assigned to the onset. Single or complex codas were then adjoined to the vowel. When the selected word had more than two syllables, unsyllabified segments

in the second or the remaining vowel were added to the preceding vowel as coda. Specifically, an R script was created to count the total number of vowels in a given entry word. A specific list of vowel types was defined at the beginning of the script. When there was only one vowel, all the consonants before the vowel were assigned to the onset of the syllable, and the remaining consonants after the vowel were assigned to the coda. If there were more than two vowels, all the consonants before the vowel were assigned to the onset of the syllable, and the search pointer was moved to the following vowel position. Then, maximal onsets between the previous and newly selected vowel, based on Duanmu's (2002:13) list of 56 onset clusters, were assigned to the current syllable, and the remaining consonants were added to the coda of the previous syllable. The procedure looped through to the final syllable. All the remaining onset consonants were assigned to the final vowel and, subsequently, all the remaining consonants were assigned to the coda slots of that final syllable. The author personally checked some samples of the syllabified outputs to avoid any inappropriate assignments before the final analysis.

### 2.3. Analysis Method

Syllabified words were processed to find the phoneme distribution using R (2016). The distribution of individual consonants and vowels was analyzed using an internal function "table" in R. Then, those consonants were grouped according to place and manner categories. The frequency distributions of the syllable numbers of the entry words were collected. The last onset and first coda consonants within each syllable were compared to find any dissimilarity between them. The immediately adjacent vowels of words with more than two syllables were also compared to examine any further dissimilarity between adjacent syllables.

## 3. Results and Discussion

### 3.1. Phoneme Distribution of Vowels and Consonants

The total number of syllables from the dictionary was 286773. That number matches the total number of vowels, while the number of consonants was 446803. The frequency ratio of vowels to consonants is approximately 4:6. Thus, one can say that English words generally contain more consonants than vowels. <Figure 1> illustrates the frequency distribution of vowels in the dictionary.

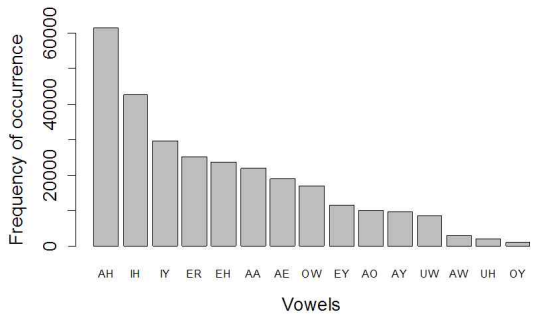


**Figure 1.** Frequency distribution of vowels in the CMU dictionary

<Figure 1> shows that the most frequently used vowel is AH with 61598 occurrences, followed by the vowel IH with 42676 occurrences. The vowel OY is recorded at the lowest frequency with 1106 occurrences. Generally, monophthongs are more prevalent than diphthongs. The percentage of front vowels is 44.2%, and that of back vowels is 55.8%. Those figures are quite comparable to the results obtained from the Buckeye Corpus (Yang, 2012), whose distribution was 48.8% for front vowels and 51.1% for back vowels.

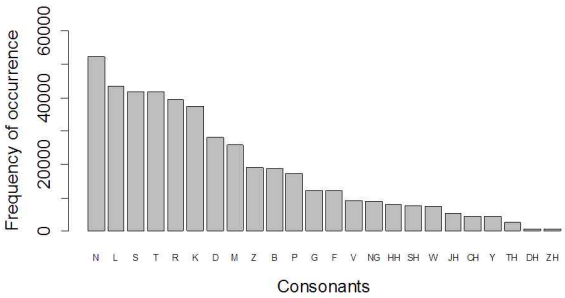


**Figure 2.** Frequency distribution of consonants in the CMU dictionary

<Figure 2> shows the distribution of consonants in the dictionary. The most frequent consonant is N with 52183 occurrences, which covers almost 11.7% of all consonants, followed by L with 43475 occurrences (9.7%). The least frequent consonant is ZH with 471 occurrences. The total number of consonants in the onset position is 294497, while 152306 consonants appear in the coda position. Approximately twice as many consonants appear in the onset position. The most frequently used single consonants in the onset position are L (20461) followed by T (18553), M (18154) and K (18110). The most frequently used double consonant clusters in the onset position are ST (5037) followed by SK (2575) and TR (2409). Triple consonant clusters, such as STR (865), SKR (215), and SPR (142), have relatively fewer occurrences. Those single consonants most frequently in the coda position are N (27701) followed by L (12209), K (8123) and NG (7805). The most frequent double consonant clusters in the coda position are NZ (2424) followed by TS (1613) and ST (1559). NTS (533), STS (368), and NDZ (193) occur less frequently. These results are slightly different from those obtained by Kessler and Treiman (1997), who analyzed 2001 entry words. They reported that /d/ was found in the onset more than other anterior coronals, but in the current study, D occurs 15921 times, much less frequently than anterior coronals such as /L, T, S/. In addition, /Z/ and /TH/ occur less often, (4834 and 1351, respectively). Thus, there may be some variations in results depending on the size of a given database.

When all the consonants are categorized by the manner and place as shown in <Table 1>, stops account for 34.6% of the total frequency of consonants, followed by fricatives (22.6%) and nasals (19.4%) in the frequency ranking. They form a distribution pattern similar to that found by Yang (2012), who reported that stops accounted for 28.1% of the Buckeye Corpus, followed by fricatives (27.1%) and nasals (20.5%).

**Table 1.** Phonetic manner categories and frequencies of English consonants in the CMU dictionary

| Manner | Frequency | % |
|---|---|---|
| stops | 154762 | 34.6 |
| fricatives | 101145 | 22.6 |
| affricates | 9572 | 2.1 |
| nasals | 86738 | 19.4 |
| laterals | 43475 | 9.7 |
| approximants | 51111 | 11.4 |
| Total | 446803 | 100.0 |

**Table 2.** Phonetic place categories and frequencies of English consonants in the CMU dictionary

| Place | Frequency | % |
|---|---|---|
| bilabial | 69026 | 15.4 |
| labiodental | 21256 | 4.8 |
| dental | 3018 | 0.7 |
| alveolar | 265380 | 59.4 |
| postalveolar | 17687 | 4.0 |
| palatal | 4232 | 0.9 |
| velar | 58262 | 13.0 |
| glottal | 7942 | 1.8 |
| Total | 446803 | 100.0 |

<Table 2> lists the percentage distribution of consonants categorized by place. The most favored place in the vocal tract was the alveolar region, specifically the alveolars (59.4%) followed by the bilabials (15.5%) and the velars (13.0%). Again, those ratios are almost comparable to the results from the Buckeye Corpus (Yang, 2012). In that study, the observed ratios were reported as 55.8%, 15.8%, and 9.5% for alveolars, bilabials, and velars, respectively.

### 3.2. Analysis of Syllable Structure

The frequency of occurrence of the number of syllables in the CMU dictionary is given in <Figure 3>. Two-syllable words are most frequent, followed by three- and one-syllable words. The author understands that two-syllable words may be one of the best options to represent different expressions for naming more objects or actions in daily use with lower memory load. If the number of syllables becomes greater, then people may have to spend more time memorizing words in their learning stage, and processing the decoding of a message from a counterpart who produces words with many more syllables may also demand more time. Two-syllable words constitute 40.7% of the total number of entry words. The total number of one-, two- and three-syllable words constitute 92.7% of all words. The percentage decreases as the number of the syllables increases. Human beings may have difficulty decoding the meanings for words with more than five syllables. Perceptual experiments to compare memory loads for words with shorter or longer syllables may help us understand why the current distribution patterns prevail in the dictionary.

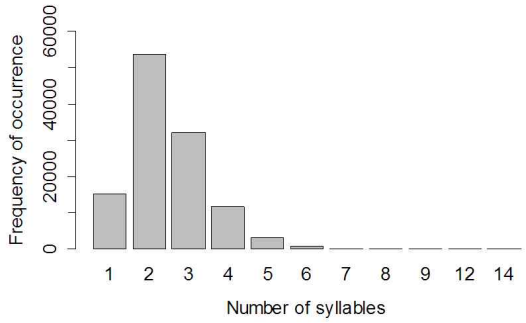


**Figure 3.** Frequency of occurrence according to the number of syllables in the CMU dictionary

As Kessler & Treiman (1997) described tendencies toward onset-coda dissimilarity in the 2001 English CVC words, the author examined the consonants and vowels of the adjacent syllables to check whether they are placed to favor discords in terms of both the manner and place of articulation. For a detailed analysis of onset or coda distributions, all word data are syllabified and assigned to a matrix with three columns of onset, peak, and coda. The total number of syllables is 286773, as described in the previous section on phoneme distribution. Then, the onset consonants are compared with the coda consonants. <Table 3> lists the frequency distribution of the syllable comparison.

**Table 3.** A comparison table of onset (O) and coda (C) in the CVC syllable structure. = means the same, while ≠ means different. # denotes no consonant in either the onset or coda.

| Category | Frequency | % |
|---|---|---|
| O≠C | 108299 | 37.8 |
| O≠# | 144506 | 50.4 |
| #≠C | 14715 | 5.1 |
| O=C | 2803 | 1.0 |
| #=# | 16450 | 5.7 |

The majority of syllables favor discord between onset and coda. The table clearly shows that 93.3% of the syllable structures exhibit discord between the onset and coda consonants. Only a small number of syllables have either the same consonants or null spaces between the onset and coda. The interpretation advanced in this study is that the discord may facilitate the perception of such words for listeners. Such discord may mean more effort to produce different sounds but less effort to perceive them. In other words, speakers do not have to pay too much attention to the last components of the syllable to deliver their thoughts clearly to listeners. In addition, the majority (50.4%) of syllable patterns consist of O≠#, followed by O≠C at 37.8%. When the syllables with onsets are summed, the total is 88.2%. Together, the syllables with codas amount to 43.7%. Further analysis of the first and second syllables reveals that 59878 first syllables are categorized as O≠# type followed by 51869 second syllables, the total accounting for

39% of all syllables. Moreover, O≠C type in the first and second syllables constitutes 28.7% of all syllables. Thus, one can conclude that the general pattern of English syllables favors onsets more than codas. This finding may make sense when we consider the importance of onsets in the semantic decision of given words in daily conversation. As soon as listeners hear the onset and the following vowel from speakers, they can easily guess the words even without listening to any consonant of the coda.

In addition to the previous consonantal analysis, words with more than two syllables are chosen to check whether the two adjacent vowels are the same or different. <Table 4> lists the frequency distribution of adjacent vowel comparisons. Out of 170070 vowels, 91.6% show discord between adjacent vowels. The percentage decreases until the 4th vowels and then increases again. Thus, we can conclude that English words favor discord in both adjacent consonants and vowels. Readers may note that there is a limitation of the current study based on the author's inference and syllabification procedure.

**Table 4.** A comparison table of adjacent vowels. V1:V2 indicates that the first vowel and the second vowel are being compared. "different" indicates that the two vowels are different, while "same" indicates that they are the same.

| Comparison | different | | same | | Total |
|---|---|---|---|---|---|
| | frequency | % | frequency | % | |
| V1:V2 | 93936 | 92.5 | 7620 | 7.5 | 101556 |
| V2:V3 | 43418 | 90.8 | 4420 | 9.2 | 47838 |
| V3:V4 | 14029 | 89.3 | 1681 | 10.7 | 15710 |
| V4:V5 | 3672 | 89.3 | 439 | 10.7 | 4111 |
| V5:V6 | 755 | 91.6 | 69 | 8.4 | 824 |
| V6:V7-V13:V14 | 30 | 96.8 | 1 | 3.2 | 31 |
| Sum | 155840 | 91.6 | 14230 | 8.4 | 170070 |

## 4. Summary and Conclusion

This study explored the phoneme distribution and syllable structure of entry words in the CMU English Pronouncing Dictionary to more deeply understand English words and to provide phoneticians and linguists with fundamental phonetic data on English words. The entry words in the dictionary file were syllabified using an R script and examined to obtain the following results.

First, English words tend to contain more consonants than vowels. In addition, monophthongs occur much more frequently than diphthongs. The proportion of front vowels is 44.2%, while that of back vowels is 55.8%. AH is listed as the most frequently used vowel, while N is listed as the most frequent consonant. When all consonants were categorized by manner and place, the distribution indicated the frequency order of stops, fricatives, and nasals according to manner and that of alveolars, bilabials and velars according to place. Those are clearly comparable to the results obtained from the Buckeye Corpus (Yang, 2012).

Second, in the analysis of syllable structure, two-syllable words were most favored, followed by three- and one-syllable words. Of the words in the dictionary, 92.7% consisted of one, two or three syllables. The results may be related to human memory or decoding time.

Third, English words tend to pursue discord both between onset and coda consonants and between adjacent vowels. Dissimilarity between the last onset and the first coda was shown in 93.3% of the syllables, while 91.6% of the adjacent vowels were different.

From the results above, the author concludes that syllabic analysis of a large database of phonetic symbols in a dictionary may lead to a deeper understanding of English word structures and components. Further studies of perceptual experiments on short or long English words are desirable to determine whether either working memory or decoding time is a main factor favoring a shorter syllable structure in English.

## References


Berg, T. (1994). The sensitivity of phonological rimes to phonetic length. *Arbeiten aus Anglistik und Amerikanistik*, 19, 63-81.

Borowsky, T. (1989). Structure preservation and the syllable coda in English. *Natural Language and Linguistic Theory*, 7, 145-166.

Cable, S. (2013). Syllables and phonotactics. Retrieved from http://people.umass.edu/scable/LING201-SP13/Slides-Handouts/Syllables-Phonotactics.pdf on March 1, 2016.

Crystal, T. H. & House, A. S. (1988). The duration of American-English stop consonants: An overview. *Journal of Phonetics*, 16, 285-294.

Davis, S. (1985). *Topics in syllable geometry.* Ph.D. Dissertation, University of Arizona, Tucson.

Duanmu, S. (2002). Two theories of onset clusters. *Chinese Phonology*, 11, 97-120.

Flexner, B. (1987). *The Random House dictionary of the English language (2nd edition).* New York: Random House.

Goldsmith, J. A. (1990). *Autosegmental and metrical phonology.* Oxford: Blackwell.

Harley, H. (2006). *English words: A linguistic introduction.* Oxford: Blackwell.

Jackson, H. (1980). *Analyzing English: An introduction to descriptive linguistics.* Oxford: Pergamon Press.

Kessler, B. & Treiman, R. (1997). Syllable structure and the distribution of phonemes in English syllables. *Journal of Memory and Language*, 37, 295-311.

McMahon, A. (2002). *An introduction to English phonology.* New York: Oxford University Press.

Noyer, R. (2016). Transcription of English syllable structure. Retrieved from http://www.ling.upenn.edu/~rnoyer/courses/103/Transcription.pdf on March 1, 2016.

R. Core Team. (2016). R: A language and environment for statistical computing. Retrieved from https://www.r-project.org/ [R Foundation for Statistical Computing, Vienna, Austria] on March 1, 2016.

Rubach, J. & Booij, G. (1990). Syllable structure assignment in Polish. *Phonology*, 7, 121-158.

Williamson, G. (2014). Syllables and clusters. Retrieved from http://www.sltinfo.com/syllables-and-clusters/ on March 1, 2016.

Yang, Byunggon. (2012). Reduction and frequency analyses of vowels and consonants in the Buckeye Speech Corpus. *Phonetics and Speech Sciences*, 4(3), 75-83.



● **Byunggon Yang**
English Education Dept.
Pusan National University
30 Changjundong, Keumjunggu,

Pusan, Korea
Tel: 051-510-2619
Email: bgyang@pusan.ac.kr
Homepage: http://fonetiks.info/bgyang
Fields of interest: Phonetics, Phonology